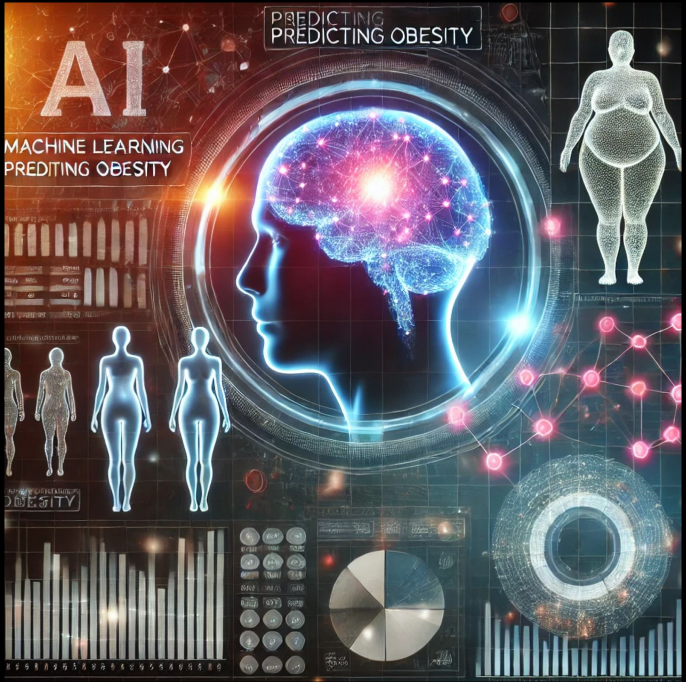
February 20th 2025



Predicting Obesity with Machine Learning

Elham Rahimi

Introduction

- Obesity is a global health concern with many medical implications
- Early predictions can help prevent obesity
- Machine learning lets us use data to predict obesity risk based on lifestyle and eating habits.

Project Objectives

- To develop a machine learning model to predict the likelihood of a person becoming obese
- Identify factors that contribute to the likelihood of a person becoming obese
- With this model; meal planners, diet centers and gyms can use this model to help their customers.

Usage

Meal Planning Services

Diet Centers

Gym

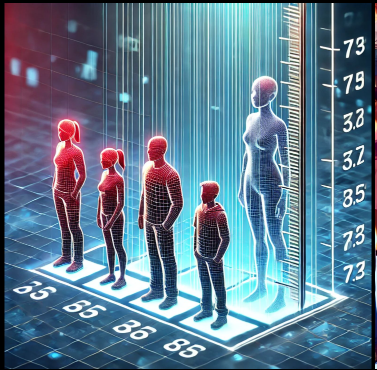
Doctors

Dataset

- Kochar, J. P. (n.d.). *Obesity Risk Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/jpkochar/obesity-risk-dataset

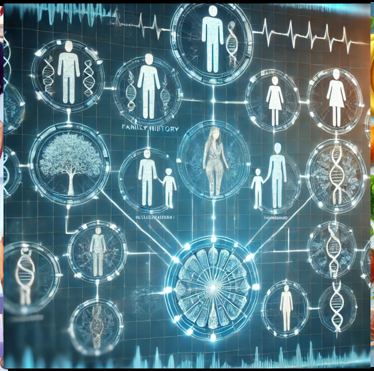
The features in the dataset include:

Anthropometric Features



Demographic Features

Genetic & Family History Features



Dietary Habits & Nutrition Features

Lifestyle Features

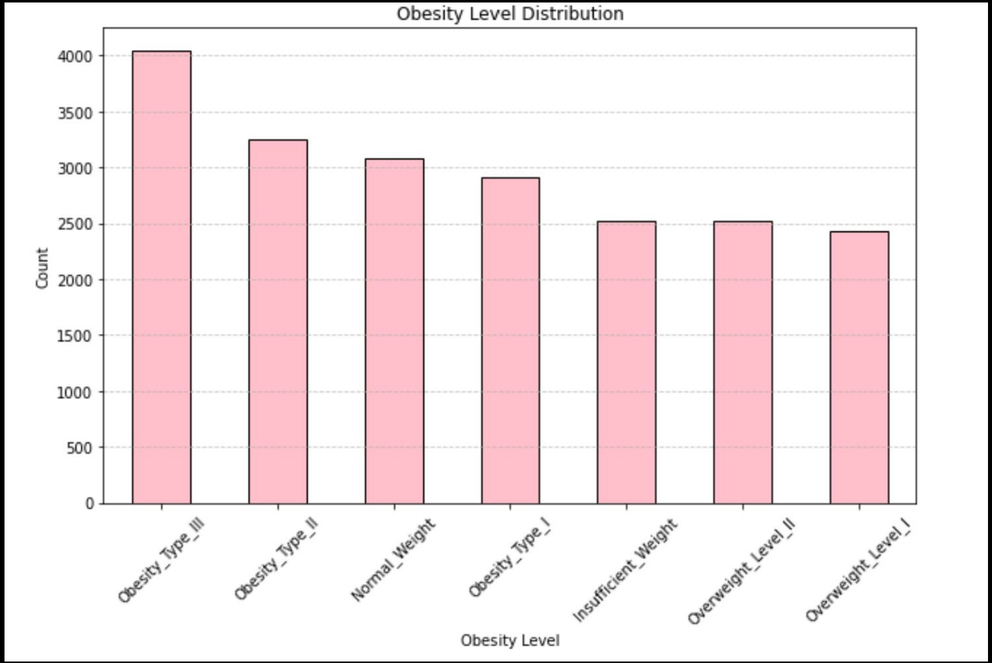
Method:

- Handling missing data and ensuring data quality before model training.
- Explore patterns, distributions and relationships between features and target variable.
- Visualize key features.
- Model train using classification algorithms.
- Model evaluation to determine accuracy.

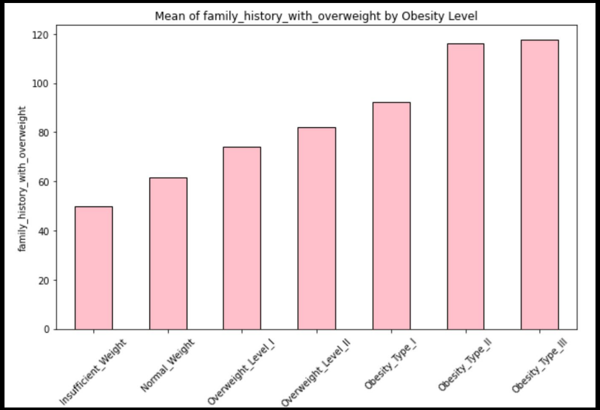
Exploratory data analysis

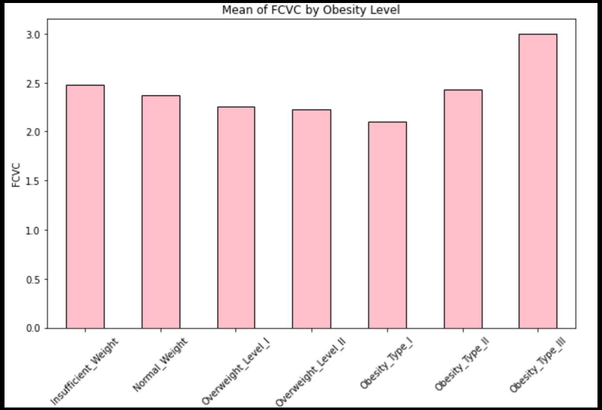
Distribution of Obesity Level by BMI

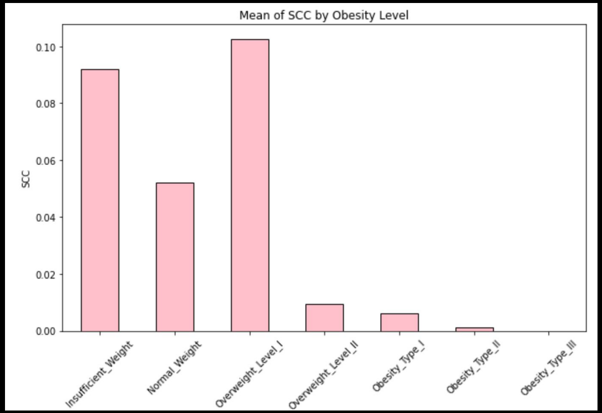
Obesity Level	Estimated BMI
Insufficient Weight	〈 18.5
Normal Weight	18.5 – 24.9
Overweight Level I	25 – 29.9
Overweight Level II	30 – 34.9
Obesity Type I	35 – 39.9
Obesity Type II	40 – 44.9
Obesity Type III	〉 45



Interesting relationships to explore





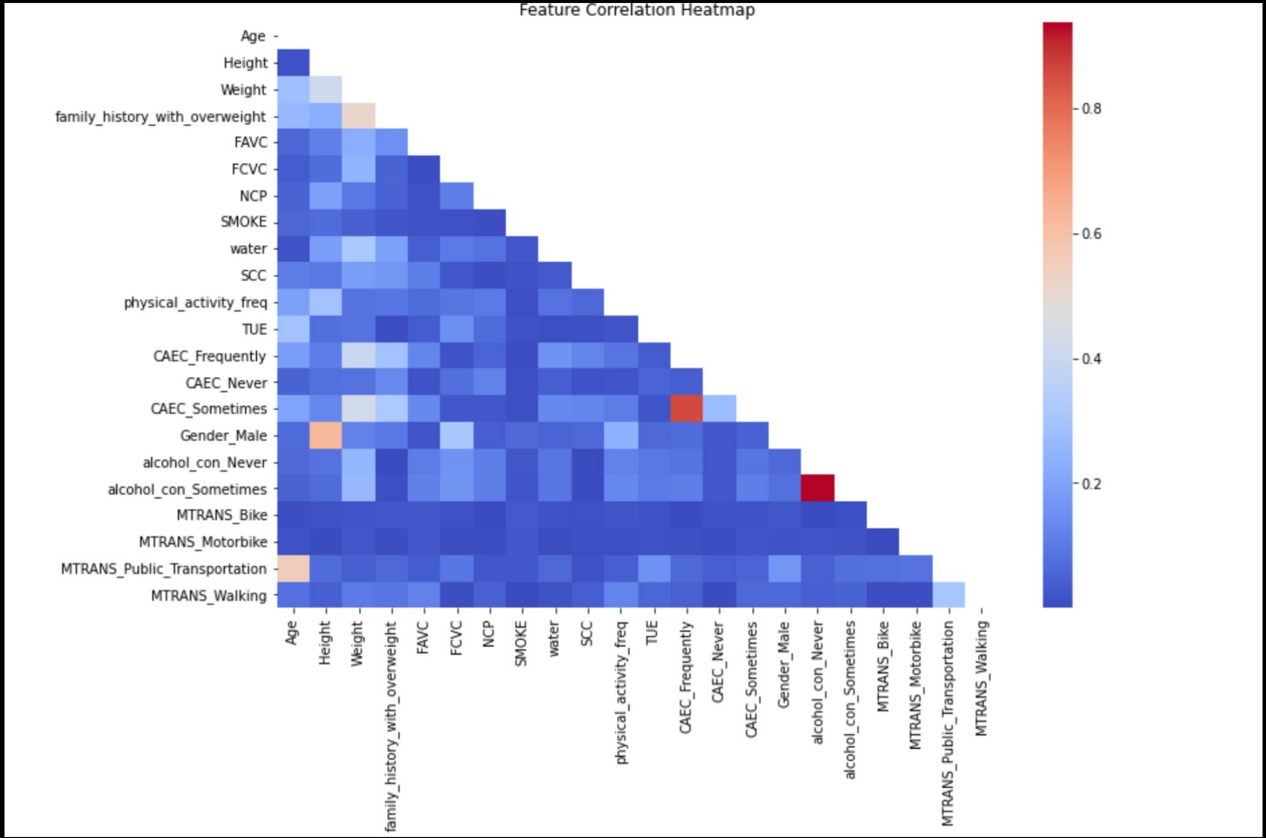


Individuals with obese family members have a higher likelihood of becoming obese themselves.

FCVC (Frequency of consumption of vegetables) was randomly distributed, but Obesity Type III consumed the most vegetables. This suggests a limit to vegetable impact on weight, as some are high in calories.

For SCC, insufficient weight and Overweight Level I consume the most caloric beverages, while Obesity Type II and III consume the least, possibly due to health awareness or dietary changes.

Preventing overfitting by checking for multicollinearity



What to do next:

- Try to balance the classes in a way.
- Remove weak features such as further reduce Multicollinearity.
- Assume the relationships aren’t linear and change the model. Next step: decision trees.